

Computing Constrained Cramér-Rao Bounds

Paul Tune, *Member, IEEE*

Abstract—We revisit the problem of computing submatrices of the Cramér-Rao bound (CRB), which lower bounds the variance of any unbiased estimator of a vector parameter θ . We explore iterative methods that avoid direct inversion of the Fisher information matrix, which can be computationally expensive when the dimension of θ is large. The computation of the bound is related to the quadratic matrix program, where there are highly efficient methods for solving it. We present several methods, and show that algorithms in prior work are special instances of existing optimization algorithms. Some of these methods converge to the bound monotonically, but in particular, algorithms converging non-monotonically are much faster. We then extend the work to encompass the computation of the CRB when the Fisher information matrix is singular and when the parameter θ is subject to constraints. As an application, we consider the design of a data streaming algorithm for network measurement.

Index Terms—Cramér-Rao bound, Fisher information, matrix functions, optimization, quadratic matrix program.

I. INTRODUCTION

The Cramér-Rao bound (CRB) [11] is important in quantifying the best achievable covariance bound on unbiased parameter estimation of n parameters θ . Under mild regularity conditions, the CRB is asymptotically achievable by the maximum likelihood estimator. The computation of the CRB is motivated by its importance in various engineering disciplines: medical imaging [6], blind system identification [20], and many others.

A related quantity is the Fisher information matrix (FIM), whose inverse is the CRB. Unfortunately, direct inversion techniques are known for their high complexity in space ($O(n^2)$ bytes of storage) and time ($O(n^3)$ floating point operations or flops). Often, one is just interested in a portion of the covariance matrix. In medical imaging applications, for example, only a small region is of importance, which is related to the location of a tumor or lesion. In this instance, computing the full inverse of the FIM becomes especially tedious and intractable when the number of parameters is large. In some applications, the FIM itself is singular, and the resulting Moore-Penrose pseudoinverse computation is even more computationally demanding. Avoiding the additional overhead incurred from direct inversion or other forms of matrix decompositions (Cholesky, QR, LU decompositions, for example) becomes a strong motivation.

Prior work [7], [8] proves the tremendous savings in memory and computation by presenting several recursive algorithms computing only submatrices of the CRB. Hero and Fessler [7] developed algorithms based on matrix splitting techniques, and statistical insight from the Expectation-

Maximization (EM) algorithm. Only $O(n^2)$ flops are required per iteration, which is advantageous if convergence is rapid, and the algorithms produce successive approximations that converge monotonically to the CRB. Exponential convergence was reported, resulting in computational savings, with the asymptotic rate of convergence governed by the relationship between the FIM of the observation space and the complete data space. This seminal work was further extended in [8], where better choices of preconditioning matrices led to much faster convergence rates. Furthermore, if the requirement of monotonic convergence of the iterates to the bound is dispensed with, there exists several algorithms with even faster convergence rates. The work also presents a way of approximating the inverse of singular FIMs.

In this paper, we show that the algorithms proposed in prior work are special instances of a more general framework related to solving a *quadratic matrix program* [1], a generalization of the well-known *quadratic program*, a convex optimization problem [2]. The reformulation provides a framework to develop methods for fast computation of the CRB, and explore various computational trade-offs. Consequently, the vast literature in convex optimization can be exploited. Our formulation enables us to extend to the cases when the parameters are constrained [4], [18] and when the Fisher information matrix is singular, with ease. The work done here may be of independent interest to other areas when a similar motivation is required. We then apply these methods on an application related to the design of a specific data streaming algorithm for measuring flows through a router. By doing so, we are able to compare the performance of several constrained optimization methods.

We denote all vectors and matrices with lower case and upper case bold letters respectively. Random variables are italicized upper case letters. Sets are denoted with upper case calligraphic font. We work entirely in the real space \mathbb{R} . \mathbb{S}_{++}^n and \mathbb{S}_+^n denote the set of real-valued, symmetric positive definite and positive semidefinite matrices of size n . The matrix $\text{diag}(\mathbf{x})$ is a diagonal matrix with elements of \mathbf{x} on its diagonals. $\text{tr}(\mathbf{A})$ and $\text{rank}(\mathbf{A})$ denote the trace and rank of a matrix \mathbf{A} respectively. The eigenvalues of \mathbf{A} are denoted by $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \dots \geq \lambda_n(\mathbf{A})$, ordered from maximum to minimum. Vector \mathbf{e}_i denotes the i -th canonical Euclidean basis in \mathbb{R}^n . $\|\mathbf{x}\|_2$ and $\|\mathbf{X}\|_F$ denotes the Euclidean and Frobenius norm of vector \mathbf{x} and matrix \mathbf{X} respectively. Other notation will be defined when needed.

II. PRELIMINARIES

A. Fisher information

Let the real, non-random parameter vector be denoted by $\theta = [\theta_1, \theta_2, \dots, \theta_n]^T$. The parameter $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^n$ is an open set. Let $\{P_\theta\}_{\theta \in \Theta}$ be a family of probability

The author is with the School of Mathematical Sciences, The University of Adelaide, Australia (Email: paul.tune@adelaide.edu.au). Technical report version, TR01-2012.

measures for a certain random variable \mathbf{Y} taking values in set \mathcal{Y} . Assume that P_θ is absolutely continuous with respect to a dominating measure μ for each $\theta \in \Theta$. Thus, for each θ there exists a density function $f(\mathbf{y}; \theta) = dP_\theta/d\mu$ for \mathbf{Y} . We define the expectation $\mathbb{E}_\theta[\mathbf{Y}] = \int \mathbf{y} dP_\theta$ whenever $\int |\mathbf{y}| dP_\theta$ is finite.

We assume that the family of densities $\{f_\mathbf{Y}(\mathbf{y}; \theta)\}_{\theta \in \Theta}$ is regular, i.e. satisfying the following three conditions: (1) $f_\mathbf{Y}(\mathbf{y}; \theta)$ is continuous on Θ for μ -almost all \mathbf{y} , (2) the log-likelihood $\log f_\mathbf{Y}(\mathbf{y}; \theta)$ is mean-square differentiable in θ , and (3) $\nabla_\theta \log f_\mathbf{Y}(\mathbf{y}; \theta)$ is mean-square continuous in θ . These conditions ensure the existence of the FIM

$$\mathbf{J}_\theta := \mathbb{E}_\theta[\nabla_\theta \log f_\mathbf{Y}(\mathbf{y}; \theta)][\nabla_\theta^T \log f_\mathbf{Y}(\mathbf{y}; \theta)], \quad (1)$$

which is an $n \times n$ positive semidefinite matrix and is finite. With the assumption of the existence, continuity in θ and absolute integrability in \mathbf{Y} of the mixed partial differential operators $(\partial^2/\partial\theta_i\partial\theta_j)f_\mathbf{Y}(\mathbf{y}; \theta)$, $i, j = 1, 2, \dots, n$, the FIM becomes equivalent to the Hessian of the mean of the curvature of $\log f_\mathbf{Y}(\mathbf{y}; \theta)$, $\mathbf{J}_\theta = -\mathbb{E}_\theta \nabla_\theta^2 \log f_\mathbf{Y}(\mathbf{y}; \theta)$.

B. Cramér-Rao Bound

The importance of the Fisher information is its relation to the Cramér-Rao bound (CRB). The CRB is a lower bound on the covariance matrix of any unbiased estimator of the parameter θ . For any unbiased estimator $\hat{\theta}(\mathbf{y})$ on observations \mathbf{y} , the relation is given by

$$\mathbb{E}[(\hat{\theta}(\mathbf{y}) - \theta)(\hat{\theta}(\mathbf{y}) - \theta)^T] \geq \mathbf{J}_\theta^{-1}. \quad (2)$$

In principle, it is possible to compute submatrices of the CRB by partitioning the Fisher information matrix into blocks, and then apply the matrix inversion lemma [5]. As reported in [7], methods such as sequential partitioning [9], Cholesky and Gaussian elimination require $O(n^3)$ flops. These methods have a high number of flops even if we are concerned with a small submatrix, for e.g. the covariance of just $m \ll n$ parameters, motivating our work.

III. FORMULATION AND ALGORITHMS

As a start, we assume a nonsingular Fisher information matrix \mathbf{J}_θ . We now consider the optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \mathbf{x}^T \mathbf{J}_\theta \mathbf{x} - \mathbf{b}^T \mathbf{x}, \quad (3)$$

an example of a quadratic program [2]. Let $F(\mathbf{x}) := \frac{1}{2} \mathbf{x}^T \mathbf{J}_\theta \mathbf{x} - \mathbf{b}^T \mathbf{x}$. In this case, the optimization problem is strictly convex and possesses a unique optimal. The unique optimal solution to this problem is $\mathbf{x}^* = \mathbf{J}_\theta^{-1} \mathbf{b}$. The generalization of the above is the *quadratic matrix program*,

$$\min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \frac{1}{2} \text{tr}(\mathbf{X}^T \mathbf{J}_\theta \mathbf{X}) - \text{tr}(\mathbf{B}^T \mathbf{X}). \quad (4)$$

Any feasible solution to (4) is a valid lower bound on the covariance of the unbiased estimate of θ , and the tightest lower bound (the CRB) is the global minimum to (4) [17]. Matrix \mathbf{B} focuses the computation on a submatrix of the CRB. The special case $\mathbf{B} = \mathbf{I}_n$ is equivalent to performing the full inverse of \mathbf{J}_θ , while setting $\mathbf{B} = \mathbf{e}_k$ enables computation of

the CRB of just a single θ_k , $k \in \{1, 2, \dots, n\}$. The optimal solution to this problem is $\mathbf{X}^* = \mathbf{J}_\theta^{-1} \mathbf{B}$ (see Appendix). Based on this, if an algorithm searches for the minimum of the optimization problems (3) or (4), it effectively computes $\mathbf{b}^T \mathbf{J}_\theta^{-1} \mathbf{b}$ and $\mathbf{B}^T \mathbf{J}_\theta^{-1} \mathbf{B}$ respectively, essentially computing the CRB.

A. Majorization-Minimization (MM) methods

The optimization problems above can be solved via the majorization-minimization (MM) method, which generalizes the Expectation-Maximization (EM) method (see [10] for a tutorial). Thus, the algorithm in [7] is a special instance of MM. We show that the recursive bounds of [7] are just the consequence of a special choice.

We first consider the vector case. Define

$$G(\mathbf{x}; \mathbf{x}^{(k)}) := \frac{1}{2} \mathbf{x}^T \mathbf{J}_\theta \mathbf{x} - \mathbf{b}^T \mathbf{x} + Q(\mathbf{x}; \mathbf{x}^{(k)}). \quad (5)$$

The function $Q(\mathbf{x}; \mathbf{x}^{(k)})$ must be chosen so that $G(\mathbf{x}; \mathbf{x}^{(k)})$ majorizes $F(\mathbf{x})$. There are two properties to satisfy: (1) $G(\mathbf{x}; \mathbf{x}^{(k)}) \geq F(\mathbf{x})$ for all \mathbf{x} , and (2) $G(\mathbf{x}^{(k)}; \mathbf{x}^{(k)}) = F(\mathbf{x}^{(k)})$. These requirements ensure that $G(\mathbf{x}; \mathbf{x}^{(k)})$ lies above the surface of $F(\mathbf{x})$ and is tangent at the point $\mathbf{x} = \mathbf{x}^{(k)}$ [10]. The function $G(\mathbf{x}; \mathbf{x}^{(k)})$ is referred to as a *surrogate function*. By these properties, MM-based algorithms converge to the CRB monotonically.

For example, suppose we have a matrix $\mathbf{P} \in \mathbb{S}_{++}^n$ and $\mathbf{P} \geq \mathbf{J}_\theta$ in the positive semidefinite sense, then

$$Q(\mathbf{x}; \mathbf{x}^{(k)}) := \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(k)})^T (\mathbf{P} - \mathbf{J}_\theta) (\mathbf{x} - \mathbf{x}^{(k)}) \quad (6)$$

is a popular choice. Minimizing (5) with the choice (6) w.r.t. \mathbf{x} results in a closed-form solution

$$\mathbf{x}^{(k+1)} = (\mathbf{I}_n - \mathbf{P}^{-1} \mathbf{J}_\theta) \mathbf{x}^{(k)} + \mathbf{P}^{-1} \mathbf{b} = \mathbf{x}^{(k)} + \mathbf{P}^{-1} (\mathbf{b} - \mathbf{J}_\theta \mathbf{x}^{(k)}). \quad (7)$$

The above is simply a *Jacobi iteration*, with a preconditioner \mathbf{P} [23]. Typically, \mathbf{P} is chosen to be diagonal or near diagonal, as this facilitates simple computation of \mathbf{P}^{-1} . Setting \mathbf{P} to the Fisher information matrix of the complete data space would yield the algorithm in [7]. For the matrix case, $G(\mathbf{X}; \mathbf{X}^{(k)}) := \frac{1}{2} \text{tr}(\mathbf{X}^T \mathbf{J}_\theta \mathbf{X} - \mathbf{B}^T \mathbf{X}) + Q(\mathbf{X}; \mathbf{X}^{(k)})$, we have the choice $Q(\mathbf{X}; \mathbf{X}^{(k)}) := \frac{1}{2} \text{tr}((\mathbf{X} - \mathbf{X}^{(k)})^T (\mathbf{P} - \mathbf{J}_\theta) (\mathbf{X} - \mathbf{X}^{(k)}))$, to obtain a Jacobi iteration.

The convergence rate for this particular choice is governed by the spectral radius $\rho(\mathbf{I}_n - \mathbf{P}^{-1} \mathbf{J}_\theta)$, which is the maximum magnitude eigenvalue of the matrix. Exponential convergence to the CRB is achieved by ensuring that $\rho(\mathbf{I}_n - \mathbf{P}^{-1} \mathbf{J}_\theta) < 1$ is as small as possible. It is for this reason $\rho(\mathbf{I}_n - \mathbf{P}^{-1} \mathbf{J}_\theta)$ is also referred to as the *root convergence factor* [23], which measures the asymptotic convergence rate.

The power of MM lies in the great freedom of choice when designing $Q(\mathbf{X}; \mathbf{X}^{(k)})$. For fast convergence, one needs to choose a $Q(\mathbf{X}; \mathbf{X}^{(k)})$ that well-approximates the quadratic objective around $\mathbf{X}^{(k)}$. Second, $Q(\mathbf{X}; \mathbf{X}^{(k)})$ is chosen in a way that it does not depend on quantities we desire, such as \mathbf{J}_θ^{-1} or is computationally expensive, for instance, a dense \mathbf{P} . These trade-offs make the algorithm design more of an art than science.

B. Gradient Descent (GD) methods

Gradient descent methods rely on minimizing the function along particular search directions. At each iteration, two crucial elements are required: the search direction $\mathbf{d}^{(k)}$, and the size of the step $\omega^{(k)}$. Algorithm 1 presents a generic outline of gradient descent methods.

Algorithm 1 Generic implementation of gradient descent

Require: ϵ , error threshold

```

1:  $\mathbf{x}^{(0)} \leftarrow \mathbf{x}_{\text{init}}$ 
2: while  $\|F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-1)})\|_2 \geq \epsilon$  do
3:    $\omega^{(k+1)} \leftarrow \arg \min_{\omega} F(\mathbf{x}^{(k)} + \omega \mathbf{d}^{(k)})$  {Exact line search}
4:    $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + \omega^{(k+1)} \mathbf{d}^{(k)}$ 
5:    $\mathbf{d}^{(k+1)} \leftarrow \Delta(\mathbf{x}^{(k+1)})$ 
6: end while

```

Gradient methods depend on the evaluation of a function $\Delta(\mathbf{x})$ which determines the search direction. For example, in classical gradient descent, this is simply gradient of $F(\mathbf{x})$ at each iterate, $\Delta(\mathbf{x}) = -\nabla_{\mathbf{x}} F(\mathbf{x}) = \mathbf{b} - \mathbf{J}_{\theta} \mathbf{x}$.

Exact line searches, however, can be computationally expensive. With a fixed choice of ω such that $\omega < 2/\lambda_n(\mathbf{J}_{\theta})$, the algorithm uses an *inexact* line search, equivalent to the *Richardson iteration* [23]. The Gauss-Seidel (GS) method performs $\Delta(\mathbf{x}^{(k)}) = -\mathbf{e}_{(k \bmod n)+1}$, $i = 1, 2, \dots, n$ [23]. Conjugate and preconditioned conjugate gradient algorithms [23] also belong to this class, where the search directions are constructed by performing a Gram-Schmidt procedure. Hence, some of the recursive algorithms presented in [8] are all instances of gradient descent methods. Unlike the algorithms presented in the previous section, these algorithms generally have *non-monotonic* convergence to the CRB. We particularly advocate preconditioned conjugate gradient algorithms for their fast convergence, requiring only simple line searches, and shown to have excellent performance in Section IV.

The Newton-Raphson descent method is unusable here, since it requires the inverse of the Hessian of the objective function, which is \mathbf{J}_{θ}^{-1} , of which we are avoiding its direct computation. For this reason, it is much better to use methods with simple line searches with low memory requirements.

The gradient search method can be applied to quadratic matrix programs. Some adaptation is needed, however, such as reformulating the problem to a suitable vectorized form (see [1]).

C. Extension to Singular Fisher Information

Suppose now \mathbf{J}_{θ} is singular. Such matrices arise in some areas, such as blind channel estimation [20] and positron emission tomography [6]. The properties of singular \mathbf{J}_{θ} were explored in [13], [17].

The approach taken in [8] was to add a perturbation to \mathbf{J}_{θ} in order to make it nonsingular, and then compute its inverse via recursive algorithms described above. This approach only yields an approximation to the CRB, with increased computational complexity. Instead, we take a completely different, more efficient, route.

Assuming $\mathbf{b} \in \text{range}(\mathbf{J}_{\theta})$, consider the optimization problem for the vector case,

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{b} - \mathbf{J}_{\theta} \mathbf{x}\|_2. \quad (8)$$

The optimization problem is convex and the solution is simply the minimum norm solution $\mathbf{x}^* = \mathbf{J}_{\theta}^+ \mathbf{b}$, where \mathbf{J}_{θ}^+ denotes the Moore-Penrose pseudoinverse [5], which is unique. Thus, we can solve for the CRB without having to resort to the approach taken by [8].

The generalization of (8) is

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \frac{1}{2} \|\mathbf{B} - \mathbf{J}_{\theta} \mathbf{X}\|_F, \quad (9)$$

assuming the column space of \mathbf{B} is in $\text{range}(\mathbf{J}_{\theta})$. Then, the minimum norm solution is $\mathbf{X}^* = \mathbf{J}_{\theta}^+ \mathbf{B}$. The optimization problems here can be solved via the MM or GD methods. As an example, using the MM method in the vector case, and choosing (6) results in $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{P}^{-1} \mathbf{J}_{\theta} (\mathbf{b} - \mathbf{J}_{\theta} \mathbf{x}^{(k)})$. A variation where $\mathbf{P} = \nu \mathbf{I}_n$ with $\nu \geq \lambda_1(\mathbf{J}_{\theta}^2)$ is the well-known *Landweber iteration* [19]. The technique also applies to problem (9).

D. Extension to Constrained Fisher Information

Certain constraints provide additional information, resulting in the reduction of estimator variance. A direct way of deriving the constrained Fisher information is to recompute the Fisher information with a new vector parameter γ such that constraints are incorporated in γ . Unfortunately, it generally requires a nontrivial alteration of the p.d.f.'s dependence on γ instead. Often, the approach is analytically intractable or numerically complex. The papers [4], [18] were motivated by this problem and provide analytic formulae to compute the constrained Fisher information matrix without reparameterization.

In the following, it is enough to assume the unconstrained Fisher information matrix $\mathbf{J}_{\theta} \in \mathbb{S}_+^n$. It has been shown that inequality constraints do not affect the CRB [4], thus, we focus entirely on equality constraints. Assume there are p consistent and nonredundant equality constraints on θ , i.e. $h(\theta) = \mathbf{0}$. Let $\mathbf{H}_{\theta} \in \mathbb{R}^{n \times p}$ denote the gradient of constraints $h(\theta)$. By the nonredundancy of the constraints, $\text{rank}(\mathbf{H}_{\theta}) = p$. Furthermore, let $\mathbf{U}_{\theta} \in \mathbb{R}^{n \times (n-p)}$ be a matrix whose column space is the orthogonal basis of the cokernel of \mathbf{H}_{θ} , i.e. $\mathbf{H}_{\theta}^T \mathbf{U}_{\theta} = \mathbf{0}$, and $\mathbf{U}_{\theta}^T \mathbf{U}_{\theta} = \mathbf{I}_{n-p}$.

Assuming that $\mathbf{U}_{\theta}^T \mathbf{J}_{\theta} \mathbf{U}_{\theta}$ is nonsingular, it has been shown that the CRB is simply [18]

$$\mathcal{I}_{\theta}^+ = \mathbf{U}_{\theta} (\mathbf{U}_{\theta}^T \mathbf{J}_{\theta} \mathbf{U}_{\theta})^{-1} \mathbf{U}_{\theta}^T. \quad (10)$$

If \mathbf{J}_{θ} is nonsingular, the above can be rewritten as $\mathcal{I}_{\theta}^+ = \mathbf{J}_{\theta}^{-1} - \mathbf{J}_{\theta}^{-1} \mathbf{H}_{\theta} (\mathbf{H}_{\theta}^T \mathbf{J}_{\theta}^{-1} \mathbf{H}_{\theta})^+ \mathbf{H}_{\theta}^T \mathbf{J}_{\theta}^{-1}$, which is equivalent to the bound derived in [4], by choosing $\mathbf{U}_{\theta} = \mathbf{I}_n - \mathbf{J}_{\theta}^{-1} \mathbf{H}_{\theta} (\mathbf{H}_{\theta}^T \mathbf{J}_{\theta}^{-1} \mathbf{H}_{\theta})^+ \mathbf{H}_{\theta}^T$.

The algorithms proposed in [7], [8] no longer apply here. It is also hard to see how the recursive algorithms can be extended to account for parameter constraints. It turns out constraints can be incorporated into our framework naturally.

The solution to the optimization problem (proof in Appendix B)

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \quad & \frac{1}{2} \text{tr}(\mathbf{X}^T \mathbf{J}_\theta \mathbf{X}) - \text{tr}(\mathbf{B}^T \mathbf{X}) \\ \text{subject to} \quad & \mathbf{H}_\theta^T \mathbf{X} = \mathbf{0}. \end{aligned} \quad (11)$$

is simply

$$\mathbf{X}^* = \mathbf{U}_\theta (\mathbf{U}_\theta^T \mathbf{J}_\theta \mathbf{U}_\theta)^{-1} \mathbf{U}_\theta^T \mathbf{B} = \mathcal{I}_\theta^+ \mathbf{B}. \quad (12)$$

From this, we have computed submatrices of the constrained CRB, extending the work in [7]. The general shift to a quadratic matrix program instead enables us to consider constraints naturally. If \mathbf{J}_θ is nonsingular, then equation (12) is equivalent to

$$\mathbf{X}^* = \mathbf{J}_\theta^{-1} \mathbf{B} - \mathbf{J}_\theta^{-1} \mathbf{H}_\theta (\mathbf{H}_\theta^T \mathbf{J}_\theta^{-1} \mathbf{H}_\theta)^+ \mathbf{H}_\theta^T \mathbf{J}_\theta^{-1} \mathbf{B}, \quad (13)$$

in agreement with the above.

The algorithms discussed previously require some modifications to account for constraints. MM methods are still applicable by ensuring constraints are built into the recursion. GD methods such as the preconditioned conjugate gradients algorithm can be adapted with constraints (for e.g. [3]). We test some of these methods below.

IV. APPLICATION

In this section, due to space limitations, we perform numerical experiments to test the efficiency of the algorithms on only one example. The application involves the optimization of a data streaming algorithm for the measurement of flows on networks, where the parameters θ are subject to constraints.

A. Data Streaming Algorithm Optimization

A *flow* is a series of packets with a common key, such as the source and destination Internet Protocol address. The *flow size* is defined as the number of packets it contains. We are interested in the flow size distribution $\theta = [\theta_1, \theta_2, \dots, \theta_n]^T$ in a measurement interval of T seconds. Each θ_k denotes the proportion of flows of size k , with n being the largest flow size. By definition, $\sum_{k=1}^n \theta_k = 1, \theta_k > 0, \forall k$ (the strict inequality of the latter constraint to ensure no bias issues arise, see [21]). The gradient of the equality constraint is $\mathbf{1}_n$.

Data streaming algorithms are used for measuring flows on core networks, since the huge volume and speed of flows imposes strict memory and processing requirements. The advantage of these algorithms is the small amount of memory required, at the expense of introducing some error when recovering flow traffic statistics.

The data streaming algorithm we consider is the flow sampling-sketcher (FSS). FSS has an array of A counters. Every incoming flow is selected with i.i.d. probability p and dropped with probability q . Sampling is performed via the use of a sampling hash function $h_s(x)$ with full range R , which acts on a flow key x , configured such that a packet is accepted if $h_s(x) \leq pR$. The deterministic nature of the hash function ensures that packets belonging to a sampled flow will be always sampled and vice versa. For packets of sampled flows, another hash function $h_c(x)$ with range A generates an index,

ensuring that the same counter is incremented by packets from the same flow. The counter with the corresponding index is incremented once per packet. Note that several flows can be mapped to the same counter, resulting in *collisions*. Once the measurement interval is over, the flow size distribution is recovered by employing an EM algorithm. FSS is practically implementable in routers. The schemes in [12], [16] are closest in spirit to FSS.

Let N_f be the total number of flows in the measurement interval and $\alpha' = pN_f/A$ denote the average number of flows in a counter. Assuming fixed A (i.e. fixed memory allocation), the latter has a direct impact on estimation quality, as α' controls the flow collisions in the counters. Sampling with low p would increase variance due to missing flows, while high p results in many flows mapping to the same counter, increasing ambiguity due to collisions. Due to the dependence between a flow size k on flows smaller than it in each counter, different optimal sampling rates p_k^* minimize the estimator variance for each θ_k . The objective is to find p_k^* for a particular target flow size k , for e.g. $k = 1$ which is especially important for detecting network attacks.

We use the Poisson approximation to compute the counter array load distribution, $\mathbf{c}_\theta = [c_0(\theta), c_1(\theta), \dots]^T$. The generating function of the load distribution is ($|s| < 1$ for convergence)

$$C^*(s; \theta) = \prod_{k=1}^n e^{\alpha' \theta_k (s^k - 1)} = e^{-\alpha'} \cdot e^{\sum_{k=1}^n \alpha' \theta_k s^k}, \quad (14)$$

essentially a convolution of n weighted Poisson mass functions (see [22]). The distribution \mathbf{c}_θ is obtained from the coefficients of the polynomial expansion of $C^*(s; \theta)$, easily computed via the Fast Fourier Transform (FFT). Examples include $c_0(\theta) = e^{-\alpha'}$, $c_1(\theta) = \alpha' \theta_1 e^{-\alpha'}$, $c_2(\theta) = (\alpha' \theta_2 + \frac{\alpha'^2 \theta_1^2}{2!}) e^{-\alpha'}$. Let $\mathbf{W}_\theta = \text{diag}(c_1^{-1}(\theta), c_2^{-1}(\theta), \dots)$. The unconstrained Fisher information is

$$\mathbf{J}_\theta(p) = q \mathbf{1}_n \mathbf{1}_n^T + p \alpha' \mathbf{G}_\theta^T \mathbf{W}_\theta \mathbf{G}_\theta, \quad (15)$$

where the matrix \mathbf{G}_θ is a quasi-Toeplitz matrix with generating sequence \mathbf{c}_θ , and $\mathbf{J}_\theta(p)$ is positive definite $\forall p$ [22]. The matrix is dense and is challenging to compute. Computing the constrained Fisher information $\mathcal{I}_\theta^+(p)$ is even more difficult (c.f. (10)). Since we only need to compute individual diagonal entries of $\mathcal{I}_\theta^+(p)$, as each k -th diagonal is the CRB of the estimator variance of θ_k , and defining

$$\begin{aligned} g(\mathbf{x}^*, p) &= \min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \mathbf{x}^T \mathbf{J}_\theta(p) \mathbf{x} - \mathbf{e}_k^T \mathbf{x} \\ \text{subject to} \quad & \mathbf{1}_n^T \mathbf{x} = 0, \end{aligned} \quad (16)$$

we can use, assuming unimodality of $g(\mathbf{x}^*, p)$ w.r.t. p , (see Appendix for justification)

$$p_k^* = \arg \max_{p \in (0, 1]} g(\mathbf{x}^*, p). \quad (17)$$

For fixed p , since problem (16) is equivalent to (11), we can use efficient optimization algorithms to solve it, avoiding full inversion. Problem (17) can be solved by a golden section search [15]. The generality of our approach allows us to, for e.g. compute $p_{[k, \ell]}^*$, the optimal sampling rate that minimizes

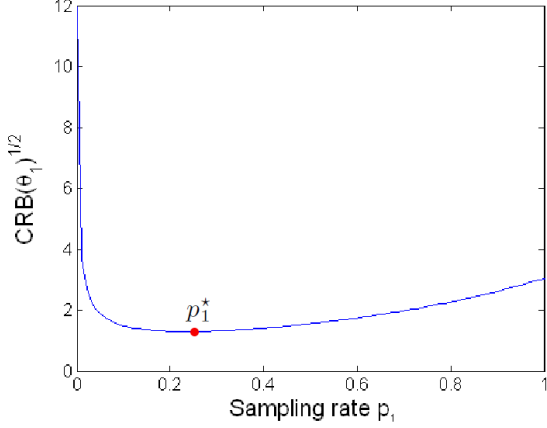


Fig. 1. Sampling rate p against the square root of CRB of θ_1 for the distribution Abilene-III, truncated at $n = 2,000$, and $\alpha = 4$. The optimal sampling rate $p_1^* = 0.2342$, denoted by the dot on the curve.

the joint variance of θ_k to θ_ℓ , achieved by replacing \mathbf{e}_k with a matrix \mathbf{B} which is zero everywhere except for the matrix $\mathbf{I}_{\ell-k+1}$ at the ℓ -th position, and use (11) in place of problem (16).

B. Numerical Results

Our focus is the computation of (16) using the various methods discussed earlier. We tested the algorithms on the important case of $k = 1$. The distribution used is a truncated version of the Abilene-III [14], truncated to $n = 2,000$ packets to satisfy the parameter constraints. Here, $\alpha = 4$, $p_1^* = 0.2342$ and tolerance for all iterative algorithms is within $\epsilon = 10^{-6}$ of the true CRB. For reference, $[\mathcal{I}_\theta^+(p_1^*)]_{11} = 1.67005$. While it is unknown if the sampling rate-CRB curve is strictly convex for all θ_k , in this case, it is (see Figure 1(a)). We omit dependence on p in the following since $p = p_1^*$.

In practice, \mathbf{c}_θ is truncated up to a sufficiently large number of terms K and computed using Fast Fourier transforms. In what follows, we assume \mathbf{c}_θ has been computed. Define $\mathbf{J}_{\theta,M}$ to be the Fisher information computed with \mathbf{c}_θ up to M terms. Then, K is chosen as the value when $\|\mathbf{J}_{\theta,K} - \mathbf{J}_{\theta,K-1}\|_F < \delta$, i.e. a preset tolerance $\delta > 0$. With the value of K terms, it takes $n^2(K+1)$ flops to construct \mathbf{J}_θ . In our case, $K = 10,000$. We assume that \mathbf{J}_θ is computed and stored upfront for all methods. In our case, this is cheaper than recomputing vector-matrix products $\mathbf{J}_\theta \mathbf{x}$, due to the complexity of \mathbf{G}_θ , at the expense of higher memory storage.

If we perform full inversion, it takes $n^3/3$ flops via Cholesky methods, followed by $3n^2$ flops to construct the term $\mathbf{J}_\theta^{-1} \mathbf{1}_n (\mathbf{1}_n^T \mathbf{J}_\theta^{-1} \mathbf{1}_n)^{-1} \mathbf{1}_n^T \mathbf{J}_\theta^{-1}$. Thus, it takes a total of $n^3/3 + 3n^2$ flops, and with $n = 2,000$, it takes 2.68 Gflops. In contrast, for recursive methods, each iteration requires $(n+1)^2$ flops, with the additional requirement to account for constraints. Depending on the method, additional operations might be required such as computing the diagonal preconditioner, which would require about $9n$ flops (see [8]). Generally, $O(n^2)$ flops per iteration are required for the following methods.

We compare two classes: Constrained Majorization-Minimization (CMM) and Constrained Preconditioned Conjugate Gradient (CPCG). For CMM, we have CMM-CF where the Fisher information of the complete data space, $\bar{\mathbf{J}}_\theta = \alpha \text{diag}(\theta_1^{-1}, \theta_2^{-1}, \dots, \theta_n^{-1})$ was used as the preconditioning matrix. CMM-DD instead uses the first order diagonally dominant matrix of \mathbf{J}_θ (see [8] for more details). The recursion step for CMM was derived using Lagrangian multipliers to account for constraints when minimizing (5). For CPCG, the preconditioner matrix used is $\bar{\mathbf{J}}_\theta$.

We also tested Gradient Projection (GP), which is the standard GD algorithm using $\bar{\mathbf{J}}_\theta$ as preconditioner, but accounts for constraints and uses exact line searches. GP, however, diverged for all iterations. Even without a preconditioner, the results remain the same. We omit its results and explain its poor performance later on.

All algorithms were initialized with the same initial point. The *breakeven threshold*, i.e. the number of iterations before all methods would lose computational advantage to a direct evaluation of the CRB is 667 iterations. Table I presents the comparison between different algorithms. The second column lists the root convergence factor of each algorithm. The root convergence factor, ρ is defined differently for each method. For CMM, refer to Section III. For CPCG, it is $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, where $\kappa = \frac{\lambda_1(\bar{\mathbf{J}}_\theta^{-1} \mathbf{J}_\theta)}{\lambda_n(\bar{\mathbf{J}}_\theta^{-1} \mathbf{J}_\theta)}$ [23]. The third and fourth columns lists the number of iterations required for the result from the algorithms to be 5% and 0.5% respectively, tolerance of the CRB. The fifth column denotes the number of iterations before the algorithm reaches within tolerance ϵ of the CRB.

While all CMM algorithms converge monotonically to the bound, they are extremely slow. The monotonic convergence of both algorithms can be seen in Figure 2. Clearly, CMM-CF has a faster convergence rate compared to CMM-DD. CPCG performs the best; however, its iterates have non-monotonic convergence. CMM and GP perform badly due to the small condition number of \mathbf{J}_θ , which is 2.73×10^6 . In particular, for GP, the projected descent steps move in a circular trajectory. Note that for all methods, the root convergence factor is a good predictor of the total number of iterations needed for convergence, but is not predictive of the number of iterations needed to be within 5% and 0.5% of the bound. Furthermore, only CPCG possesses some robustness with respect to the selection of the initial point. The other algorithms have a strong dependence on the initial point, and may have poor performance with a bad initial point choice. Finally, CPCG is the only algorithm that converges within the breakeven threshold.

We also compare CPCG against a straightforward way of evaluation using the GS method. We use GS to evaluate two quantities separately: $\mathbf{J}_\theta^{-1} \mathbf{e}_1$ and $\mathbf{J}_\theta^{-1} \mathbf{1}_n$, and then use (13) to evaluate \mathbf{x}^* . The trajectory of this method was compared with the trajectory of CPCG in Figure 3. Initialization for GS requires two initial points for evaluation of the two quantities. To ensure fairness, CPCG was initialized using the first evaluated point of GS. GS reaches to within 5% and 0.5% of the bound in 8 and 10 iterations, and requires 110 iterations for convergence. In contrast, CPCG takes 5, 6 and

Alg.	ρ	5%	0.5%	Convergence
CMM-DD	0.9998	12,935	23,549	69,026
CMM-CF	0.9986	161	407	8,487
CPCG	0.8715	5	7	48

TABLE I
ASYMPTOTIC AND FINITE CONVERGENCE PROPERTIES OF THE ITERATIVE ALGORITHMS

64 iterations for 5% and 0.5% of the bound, and convergence respectively. The reason for the slow convergence of GS is due to the oscillations occurring near the end, as this method does not perform a methodical search across the constrained space, unlike CPCG. Clearly, CPCG is far superior to this method. As seen in Figure 3, both methods converge non-monotonically to the true CRB.

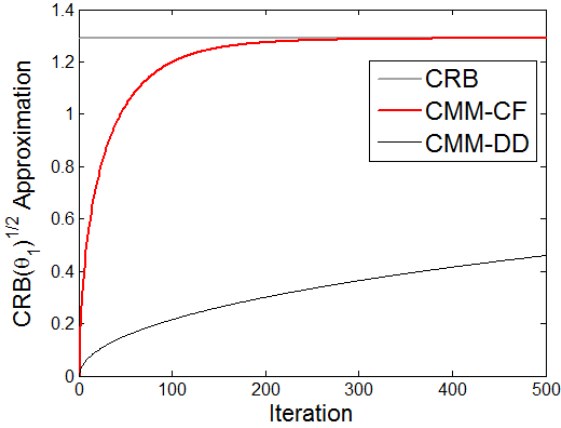


Fig. 2. Trajectory of CMM-CF and CMM-DD when computing the square root of CRB of θ_1 for the distribution Abilene-III, truncated at $n = 2,000$, and $\alpha = 4$, shown here up to 500 iterations. Tolerance $\epsilon = 10^{-6}$. CMM-CF converged in 8,487 iterations, while CMM-DD converged in 69,026 iterations. Note the monotonic convergence of both algorithms to the true CRB.

V. CONCLUSION

In this paper, we revisit the problem of computing submatrices of the CRB. We show that computation of these submatrices are related to a quadratic matrix program. Due to the properties of the FIM and the convexity of the quadratic matrix program, we can compute the submatrices with efficient algorithms from convex optimization literature. We further show how the framework here easily extends to the case when the FIM is singular, and when parameter constraints are present. We then apply the algorithms on a constrained optimization problem, showing that the computation of these bounds can be evaluated efficiently for important signal processing problems. Future work includes exploring more algorithms for evaluation that may possess faster convergence rates and testing on other constrained problems.

APPENDIX

A. Derivation of the Optimal Solution of (4)

The derivation relies on the relations $\nabla_{\mathbf{X}} \text{tr}(\mathbf{X}^T \mathbf{J}_\theta \mathbf{X}) = 2\mathbf{J}_\theta \mathbf{X}$ and $\nabla_{\mathbf{X}} \text{tr}(\mathbf{B}^T \mathbf{X}) = \mathbf{B}$. Using these relations, the

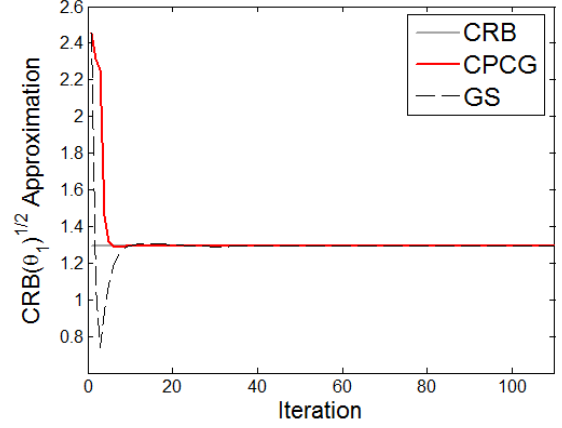


Fig. 3. Trajectory of the GS and CPCG when computing the square root of CRB of θ_1 for the distribution Abilene-III, truncated at $n = 2,000$, and $\alpha = 4$. Tolerance $\epsilon = 10^{-6}$. GS converged in 110 iterations, while CPCG converged in 64 iterations.

gradient of the problem is $\mathbf{J}_\theta \mathbf{X} - \mathbf{B}$. Setting this to 0, we obtain the optimal solution.

B. Derivation of (12)

Since $\mathbf{J}_\theta \in \mathbb{S}_+^n$, the objective is a convex function, and the constraints are linear, thus, the optimization problem remains convex. By using Lagrangian multipliers \mathbf{Z} , the optimal solution obeys $\mathbf{J}_\theta \mathbf{X} - \mathbf{B} - \mathbf{H}_\theta^T \mathbf{Z} = \mathbf{0}$ and $\mathbf{H}_\theta^T \mathbf{X} = \mathbf{0}$. This implies that the feasible solutions of \mathbf{X} has the structure $\mathbf{X} = \mathbf{U}_\theta \mathbf{Y}$, where $\mathbf{Y} \in \mathbb{R}^{(n-p) \times m}$, as solutions must lie in the cokernel of \mathbf{H}_θ and the range space of \mathbf{U}_θ . Proceeding in this fashion, we obtain $\mathbf{J}_\theta \mathbf{U}_\theta \mathbf{Y} = \mathbf{B}$. Multiplying by \mathbf{U}_θ^T on both sides, we then get $\mathbf{Y}^* = (\mathbf{U}_\theta^T \mathbf{J}_\theta \mathbf{U}_\theta)^{-1} \mathbf{U}_\theta^T \mathbf{B}$. Multiplying \mathbf{Y}^* again by \mathbf{U}_θ , we obtain the optimal solution \mathbf{X}^* . In the case of nonsingular \mathbf{J}_θ , one can choose $\mathbf{U}_\theta = \mathbf{I}_n - \mathbf{J}_\theta^{-1} \mathbf{H}_\theta (\mathbf{H}_\theta^T \mathbf{J}_\theta^{-1} \mathbf{H}_\theta)^+ \mathbf{H}_\theta^T$ as it lies in the cokernel of \mathbf{H}_θ and is orthogonal (see details in [4]). Then, \mathbf{X}^* is equivalent to (13).

C. Formulation of (17)

Consider the objective function once the inner minimization problem is solved. The objective function yields $-\frac{1}{2}[\mathcal{I}_\theta^+(p')]_{kk}$, for some particular rate sampling rate p' . Now, p_k^* is the optimal value if and only if $-\frac{1}{2}[\mathcal{I}_\theta^+(p_k^*)]_{kk} > -\frac{1}{2}[\mathcal{I}_\theta^+(p')]_{kk}$ for all other $p' \neq p_k^*$. By maximizing the objective function over p , we solve for p_k^* .

D. Derivation of the Constrained MM

We prove the result for the vector case. Similar derivation applies for the matrix case. As discussed in Section III, we use $Q(\mathbf{x}; \mathbf{x}^{(k)}) := \frac{1}{2}(\mathbf{x} - \mathbf{x}^{(k)})^T (\mathbf{P} - \mathbf{J}_\theta)(\mathbf{x} - \mathbf{x}^{(k)})$. Then, the task is to minimize

$$G(\mathbf{x}; \mathbf{x}^{(k)}) := \frac{1}{2} \mathbf{x}^T \mathbf{J}_\theta \mathbf{x} - \mathbf{b}^T \mathbf{x} + Q(\mathbf{x}; \mathbf{x}^{(k)}).$$

subject to the constraint $\mathbf{H}_\theta^T \mathbf{x} = \mathbf{0}$.

Using the method of Lagrangian multipliers [2], we construct the Lagrangian, with multipliers $\boldsymbol{\mu} \in \mathbb{R}^p$,

$$L(\mathbf{x}, \boldsymbol{\mu}; \mathbf{x}^{(k)}) = G(\mathbf{x}; \mathbf{x}^{(k)}) + \boldsymbol{\mu}^T \mathbf{H}_\theta^T \mathbf{x}. \quad (18)$$

At the optimal point, there are two equations to satisfy:

$$\begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}; \mathbf{x}^{(k)}) &= \mathbf{P}\mathbf{x} + (\mathbf{P} - \mathbf{J}_\theta)\mathbf{x}^{(k)} - \mathbf{b} + \mathbf{H}_\theta \boldsymbol{\mu} = \mathbf{0}, \\ \nabla_{\boldsymbol{\mu}} L(\mathbf{x}, \boldsymbol{\mu}; \mathbf{x}^{(k)}) &= \mathbf{H}_\theta^T \mathbf{x} = \mathbf{0}. \end{aligned}$$

Solving both equations, we have

$$\boldsymbol{\mu}^{(k+1)} = (\mathbf{H}_\theta^T \mathbf{P}^{-1} \mathbf{H}_\theta)^{-1} \mathbf{H}_\theta^T \left((\mathbf{I}_n - \mathbf{P}^{-1} \mathbf{J}_\theta) \mathbf{x}^{(k)} - \mathbf{P}^{-1} \mathbf{b} \right),$$

which exists, since we choose $\mathbf{P} \in \mathbb{S}_{++}^n$. Finally,

$$\begin{aligned} \mathbf{x}^{(k+1)} &= (\mathbf{I}_n - \mathbf{P}^{-1} \mathbf{J}_\theta) \mathbf{x}^{(k)} - \mathbf{P}^{-1} \mathbf{b} - \mathbf{H}_\theta \boldsymbol{\mu}^{(k+1)} \\ &= \mathbf{T}_\theta \left((\mathbf{I}_n - \mathbf{P}^{-1} \mathbf{J}_\theta) \mathbf{x}^{(k)} - \mathbf{P}^{-1} \mathbf{b} \right). \end{aligned}$$

where $\mathbf{T}_\theta = \mathbf{I}_n - \mathbf{H}_\theta (\mathbf{H}_\theta^T \mathbf{P}^{-1} \mathbf{H}_\theta)^{-1} \mathbf{H}_\theta^T$ is a projection operator. Note its similarity to the basic Jacobi iteration, except with the projection \mathbf{T}_θ to account for the parameter constraints.

REFERENCES

- [1] A. Beck. Quadratic matrix programming. *SIAM J. Optim.*, 17(4):1224–1238, 2007.
- [2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [3] T. F. Coleman. Linearly constrained optimization and projected preconditioned conjugate gradients. In *Proc. 5th SIAM Conf. on App. Lin. Algebra*, pages 118–122, 1994.
- [4] J. D. Gorman and A. O. Hero. Lower bounds for parametric estimation with constraints. *IEEE Trans. Info. Theory*, 36(6):1285–1301, November 1990.
- [5] D. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, 1997.
- [6] A. Hero, J. Fessler, and M. Usman. Exploring estimator bias-variance tradeoffs using the Uniform CR bound. *IEEE Trans. Sig. Proc.*, 44(8):2026–2041, August 1996.
- [7] A. O. Hero and J. Fessler. A recursive algorithm for computing Cramér-Rao-type bounds on estimator covariance. *IEEE Trans. Info. Theory*, 40(4):1205–1210, July 1994.
- [8] A. O. Hero, M. Usman, A. C. Sauve, and J. Fessler. Recursive algorithms for computing the Cramér-Rao bound. Technical Report 305, Communication and Signal Processing Laboratory, University of Michigan, Ann Arbor, November 1996.
- [9] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
- [10] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, February 1996.
- [11] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall PTR, March 1993.
- [12] A. Kumar, M. Sung, J. Xu, and J. Wang. Data streaming algorithms for efficient and accurate estimation of flow size distribution. In *Proc. of ACM SIGMETRICS 2004*, New York, June 2004.
- [13] R. C. Liu and L. D. Brown. Nonexistence of informative unbiased estimators in singular problems. *Ann. Stat.*, 21(1), 1993.
- [14] NLANR. Abilene-III Trace Data. <http://pma.nlanr.net/Special/ipls3.html>.
- [15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd edition, 2007.
- [16] B. Ribeiro, T. Ye, and D. Towsley. A resource minimalist flow size histogram estimator. In *Proc. 2008 ACM SIGCOMM Internet Measurement Conference*, pages 285–290, Vouliagmeni, Greece, October 2008.
- [17] P. Stoica and T. Marzetta. Parameter estimation problems with singular information matrices. *IEEE Trans. Sig. Proc.*, 49(1), January 2001.
- [18] P. Stoica and B. C. Ng. On the Cramér-Rao bound under parametric constraints. *IEEE Signal Processing Letters*, 5(7):177–179, July 1998.
- [19] O. N. Strand. Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind. *SIAM J. Numer. Anal.*, 11:798–825, September 1974.
- [20] J. R. Treichler. Special issue on: Blind system identification and estimation. *Proc. IEEE*, 86, October 1998.
- [21] P. Tune and D. Veitch. Fisher information in flow size distribution estimation. *IEEE Trans. Info. Theory*, 57(10):7011–7035, October 2011.
- [22] P. Tune and D. Veitch. Sampling vs Sketching: An Information Theoretic Comparison. In *IEEE Infocom 2011*, pages 2105–2113, Shanghai, China, April 10–15 2011.
- [23] D. S. Watkins. *Fundamentals of Matrix Computations*. Wiley-Interscience, 2nd edition, 2002.

